McRae, Jeremy F., Clayton, Stephen, Fitzgerald, Tomas W., Kaplanis, Joanna, Prigmore, Elena, Rajan, Diana, Sifrim, Alejandro, Aitken, Stuart, Akawi, Nadia, Alvi, Mohsan, Ambridge, Kirsty, Barrett, Daniel M., Bayzetinova, Tanya, Jones, Philip, Jones, Wendy D., King, Daniel, Krishnappa, Netravathi, Mason, Laura E., Singh, Tarjinder, Tivey, Adrian R., Ahmed, Munaza, Anjum, Uruj, Archer, Hayley, Armstrong, Ruth, Awada, Jana, Balasubramanian, Meena, Banka, Siddharth, Baralle, Diana, Barnicoat, Angela, Batstone, Paul, Baty, David, Bennett, Chris, Berg, Jonathan, Bernhard, Birgitta, Bevan, A. Paul, Bitner-Glindzicz, Maria, Blair, Edward, Blyth, Moira, Bohanna, David, Bourdon, Louise, Bourn, David, Bradley, Lisa, Brady, Angela, Brent, Simon, Brewer, Carole, Brunstrom, Kate, Bunyan, David J., Burn, John, Canham, Natalie, Castle, Bruce, Chandler, Kate, Chatzimichali, Elena, Cilliers, Deirdre, Clarke, Angus ORCID: https://orcid.org/0000-0002-1200-9286, Clasper, Susan, Clayton-Smith, Jill, Clowes, Virginia, Coates, Andrea, Cole, Trevor, Colgiu, Irina, Collins, Amanda, Collinson, Morag N., Connell, Fiona, Cooper, Nicola, Cox, Helen, Cresswell, Lara, Cross, Gareth, Crow, Yanick, D'Alessandro, Mariella, Dabir, Tabib, Davidson, Rosemarie, Davies, Sally, de Vries, Dylan, Dean, John, Deshpande, Charu, Devlin, Gemma, Dixit, Abhijit, Dobbie, Angus, Donaldson, Alan, Donnai, Dian, Donnelly, Deirdre, Donnelly, Carina, Douglas, Angela, Douzgou, Sofia, Duncan, Alexis, Eason, Jacqueline, Ellard, Sian, Ellis, Ian, Elmslie, Frances, Evans, Karenza, Everest, Sarah, Fendick, Tina, Fisher, Richard, Flinter, Frances, Foulds, Nicola, Fry, Andrew ORCID: https://orcid.org/0000-0001-9778-6924, Fryer, Alan, Gardiner, Carol, Gaunt, Lorraine, Ghali, Neeti, Gibbons, Richard, Gill, Harinder, Goodship, Judith, Goudie, David, Gray, Emma, Green, Andrew, Greene, Philip, Greenhalgh, Lynn, Gribble, Susan, Harrison, Rachel, Harrison, Lucy, Harrison, Victoria, Hawkins, Rose, He, Liu, Hellens, Stephen, Henderson, Alex, Hewitt, Sarah, Hildyard, Lucy, Hobson, Emma, Holden, Simon, Holder, Muriel, Holder, Susan, Hollingsworth, Georgina, Homfray, Tessa, Humphreys, Mervyn, Hurst, Jane, Hutton, Ben, Ingram, Stuart, Irving, Melita, Islam, Lily, Jackson, Andrew, Jarvis, Joanna, Jenkins, Lucy, Johnson, Diana, Jones, Elizabeth, Josifova, Dragana, Joss, Shelagh, Kaemba, Beckie, Kazembe, Sandra, Kelsell, Rosemary, Ke[illegible] Bronwyn, Kingston, Helen, Kini, Usha, Kinning, Esther, Kirby, Gail, Kirk, [illegible]re, Kivuva, Emma, Kraus, Alison, Kumar, Dhavendra, Kumar, V. K. Ajith, [illegible]an, Katherine, Lam, Wayne, Lampe, Anne, Langman, Caroline, Lees, Mel[illegible]a, Lim, Derek, Longman, Cheryl, Lowther, Gordon, Lynch, Sally A.,

ARTICLE

doi:10.1038/nature21062

Prevalence and architecture of *de novo* mutations in developmental disorders

Deciphering Developmental Disorders Study

The genomes of individuals with severe, undiagnosed developmental disorders are enriched in damaging *de novo* mutations (DNMs) in developmentally important genes. Here we have sequenced the exomes of 4,293 families containing individuals with developmental disorders, and meta-analysed these data with data from another 3,287 individuals with similar disorders. We show that the most important factors influencing the diagnostic yield of DNMs are the sex of the affected individual, the relatedness of their parents, whether close relatives are affected and the parental ages. We identified 94 genes enriched in damaging DNMs, including 14 that previously lacked compelling evidence of involvement in developmental disorders. We have also characterized the phenotypic diversity among these disorders. We estimate that 42% of our cohort carry pathogenic DNMs in coding sequences; approximately half of these DNMs disrupt gene function and the remainder result in altered protein function. We estimate that developmental disorders caused by DNMs have an average prevalence of 1 in 213 to 1 in 448 births, depending on parental age. Given current global demographics, this equates to almost 400,000 children born per year.

Approximately 2–5% of children are born with major congenital malformations and/or manifest severe neurodevelopmental disorders during childhood[1,2]. Although diverse factors, including gestational infection and maternal alcohol consumption, can cause such developmental disorders (DDs), damaging genetic variation in developmentally important genes makes a major contribution. Several recent studies have identified a substantial causal role for DNMs that are not present in either parent[3–16]. Despite the identification of many DDs caused by DNMs, it is generally accepted that many more disorders have not yet been discovered[15], and the overall contribution of DNMs to DDs is not known. Moreover, some pathogenic DNMs completely ablate the encoded protein, whereas others instead alter the function of the encoded protein[17]; the relative contributions of these two mechanistic classes are also not known.

We recruited 4,293 individuals to the Deciphering Developmental Disorders (DDD) study[15] via the genetics services of the UK National Health Service and the Republic of Ireland. Each of these individuals was referred with a severe undiagnosed DD and most were the only affected family member. Most individuals (81%) had been screened for large pathogenic deletions and duplications. We systematically phenotyped these individuals and sequenced the exomes of the affected individuals and their parents. Growth measurements, family history and developmental milestones were recorded, and detailed clinical phenotypes were captured using Human Phenotype Ontology (HPO) terms. Analyses of 1,133 of these individuals have been described previously[15,18]. We generated a high-sensitivity set of 8,361 candidate DNMs in the coding or splicing sequence (mean = 1.95 DNMs per proband), while removing systematic erroneous calls (Supplementary Table 1). This rate of candidate DNMs per proband is higher than those found in other studies[3–15], because we wanted to maintain high sensitivity and we could address the lower specificity through subsequent validation. We found that 1,624 genes contained two or more DNMs in unrelated individuals.

Twenty-three per cent of individuals had protein-truncating or missense DNMs that were probably pathogenic within the clinically curated set of genes robustly associated with dominant DDs[18]. We investigated factors associated with whether an individual had a probably pathogenic DNM in these curated genes (Fig. 1a, b and Extended Data Table 1). Males had a lower chance of carrying a probably pathogenic DNM ($P = 1.6 \times 10^{-4}$; odds ratio (OR) = 0.75; 95% confidence interval (CI) = 0.65–0.87), as has also been observed for autism[19]. We also observed that a greater extent of speech delay was correlated with an increased likelihood of having a pathogenic DNM ($P = 0.00115$), but no other indicators of severity relative to the rest of the cohort showed significant correlations. Individuals with other affected family members were less likely to have pathogenic DNMs (affected siblings: $P = 7.3 \times 10^{-18}$, affected parents: $P = 5.7 \times 10^{-9}$), as were individuals who were from self-declared consanguineous unions ($P = 8.0 \times 10^{-11}$). Furthermore, the total genomic extent of autozygosity (owing to parental relatedness) was negatively correlated with the likelihood of having a pathogenic DNM ($P = 1.7 \times 10^{-7}$). For every $\log_{10}$ increase in autozygous length, the probability of having a pathogenic DNM dropped by 7.5%, probably owing to increasing burden of recessive causation (Fig. 1c). Nonetheless, 6% of individuals with autozygosity equivalent to a first cousin union or closer had a plausibly pathogenic DNM, underscoring the importance of considering *de novo* causation in all families.

Paternal age has been shown to be the primary factor influencing the number of DNMs in a child[20,21], and is therefore expected to be a risk factor for pathogenic DNMs. Paternal age was only weakly associated with the probability of having a pathogenic DNM ($P = 0.016$). However, focusing on the minority of DNMs that were truncating and missense variants in known DD-associated genes limits our power to detect such an effect. Analysis of all 8,409 high-confidence exonic and intronic autosomal DNMs confirmed a strong effect of paternal age ($P = 1.4 \times 10^{-10}$, 1.53 DNMs per year, 95% CI = 1.07–2.01), and a weaker, independent, effect of maternal age ($P = 0.0019$, 0.86 DNMs per year, 95% CI = 0.32–1.40; Fig. 1d, e), as has recently been described using whole-genome analyses[22]. These genome-wide estimates were based on exome-based estimates, with a paternal effect of 0.0306 DNMs per year and a maternal effect of 0.0172 DNMs per year.

We identified genes that were significantly enriched for damaging DNMs by comparing the observed gene-wise DNM count to the expected count under a null-mutation model[23], as described

Lists of participants and their affiliations appear in the online version of the paper.

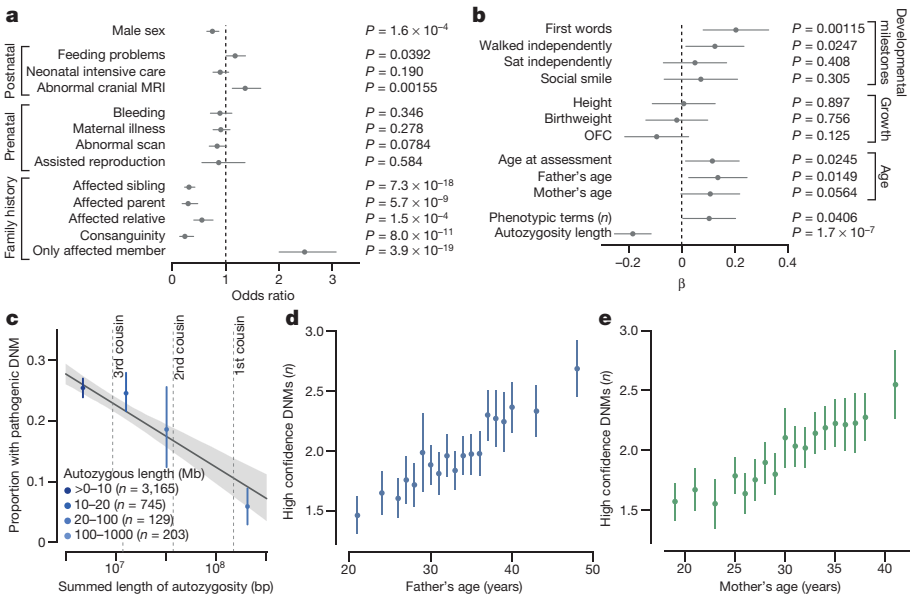


Figure 1 | Association of phenotypes with the presence of DNMs that are probably pathogenic. a, Odds ratios for binary phenotypes. Positive odds ratios are associated with increased risk of pathogenic DNMs when the phenotype is present. *P* values are given (Fisher's exact test). **b**, β coefficients from logistic regression of quantitative phenotypes versus presence of a pathogenic DNM. All phenotypes aside from length of autozygous regions were corrected for gender as a covariate. The developmental milestones (age to achieve first words, walk independently, sit independently and social smile) were log scaled before regression. The growth parameters (height, birthweight and occipitofrontal circumference (OFC)) were evaluated as absolute distance from the median. **c**, The relationship between length of autozygous regions and chance of having a pathogenic DNM. The regression line and 95% CI are plotted as the dark grey line and grey shading, respectively. The autozygosity lengths expected under different degrees of consanguineous unions are shown as vertical dashed lines. *n*, number of individuals in each autozygosity group. Blue dots and blue lines, mean ± 95% CI. **d**, Relationship between age of fathers at child birth and the number of high confidence DNMs. Error bars, 95% CI. **e**, Relationship between age of mothers at child birth and number of high confidence DNMs. Error bars, 95% CI.

previously[15]. We combined this analysis with 4,224 published DNMs in 3,287 affected individuals from thirteen exome- or genome-sequencing studies[3–14] (Supplementary Table 2) that exhibited a similar excess of DNMs as our curated set of DD-associated genes (Extended Data Fig. 1). We found 93 genes with genome-wide significance ($P < 5 \times 10^{-7}$; Fig. 2), 80 of which had previous evidence of an association with DDs (Supplementary Table 3). We have developed visual summaries of the phenotypes associated with each gene to facilitate clinical use. In addition, we created anonymised, average face images from individuals with DNMs in genome-wide significant genes (Fig. 2) from ordinary (2D) clinical photos using previously validated software[24]. These images highlight facial dysmorphologies specific to certain genes. After careful review by two experienced clinical geneticists, average face images for twelve genes were determined to be truly anonymised and of sufficient quality. To assess any increase in power to detect novel DD-associated genes, we excluded individuals with probably pathogenic variants in known DD-associated genes[15], leaving 3,158 probands from our cohort, along with 2,955 probands from the meta-analysis studies. In this subset, fourteen genes for which no statistically compelling previous evidence for DD causation was available achieved genome-wide significance: *CDK13*, *CHD4*, *CNOT3*, *CSNK2A1*, *GNAI1*, *KCNQ3*, *MSL3*, *PPM1D*, *PUF60*, *QRICH1*, *SET*, *KMT5B* (also known as *SUV420H1*), *TCF20* and *ZBTB18* ($P < 5 \times 10^{-7}$; Table 1 and Extended Data Fig. 4). The clinical features associated with these newly confirmed disorders are summarized in Extended Data Figs 2, 3 and Supplementary Information. *QRICH1* did not achieve genome-wide significance without excluding individuals with probably pathogenic variants in DD-associated genes. In addition to discovering novel DD-associated genes, we identified several new disorders linked to known DD-associated genes, but with different modes of inheritance or molecular mechanisms. We found that *USP9X* and *ZC4H2* had a genome-wide significant excess of DNMs in female probands, indicating these genes have X-linked dominant modes of inheritance in addition to a previously reported X-linked recessive mode of inheritance in males[25,26]. In addition, we found that truncating mutations in *SMC1A* were strongly associated with a novel seizure disorder ($P = 6.5 \times 10^{-19}$), whereas in-frame and/or missense mutations in *SMC1A* with dominant negative effects[27] are a known cause of Cornelia de Lange Syndrome. Individuals with truncating mutations in *SMC1A* lacked the characteristic facial dysmorphology of Cornelia de Lange Syndrome.

We then explored two approaches for integrating phenotypic data into disease gene association: statistical assessment of HPO-term similarity between individuals sharing candidate DNMs in the same gene (as previously described[28]) and phenotypic stratification based on specific clinical characteristics. Combining genetic evidence and HPO-term similarity increased the significance of some known DD-associated genes. However, significance decreased for a larger number of genes that caused severe DDs, but that are associated with non-discriminative HPO terms (Extended Data Fig. 5a). Although we did not incorporate categorical phenotypic similarity into the gene-discovery analyses described above, the systematic acquisition of phenotypic data from affected individuals within the DDD cohort enabled aggregate representations to be created for each gene that achieved genome-wide significance. We present these in the form of icon-based summaries of growth and developmental milestones (PhenIcons), heatmaps of the recurrently coded HPO terms and, where photos for at least ten children with mutations in the same gene were available, an anonymised average facial representation (Supplementary Information).

Twenty per cent of individuals had HPO terms that indicated seizures and/or epilepsy. We compared analysis within this phenotypically stratified group with gene-wise analyses of the entire cohort, to see whether it increased power to detect known seizure-associated genes (Extended Data Fig. 5b). Fifteen seizure-associated genes had genome-wide significance ($P < 5 \times 10^{-7}$) in both the seizure-only and the entire-cohort analyses. Nine seizure-associated genes had genome-wide significance ($P < 5 \times 10^{-7}$) in the entire cohort but not in the seizure subset. Of the 285 individuals with truncating or missense

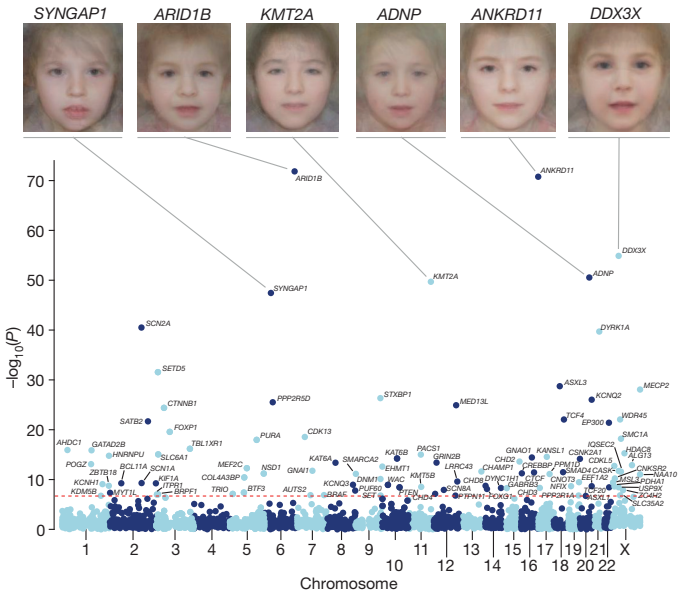


Figure 2 | Genes exceeding genome-wide significance. Manhattan plot of combined *P* values across all tested genes. The red dashed line indicates the threshold for genome-wide significance ($P < 7 \times 10^{-7}$). Genes exceeding this threshold have labelled HGNC symbols. De-identified realistic average ('composite') faces were generated using previously validated software[24] from clinical photos of individuals with DNMs in the same gene, and are shown here for the six most significantly associated genes. Confirmation of de-identification was performed by careful review by two experienced clinical geneticists. Each face was generated from clinical photos of more than ten children.

Table 1 | Genes achieving genome-wide significant statistical evidence without previous compelling evidence for association with DDs

Gene	Missense	PTV	*P* value	Test	Clustering
CDK13	10	1	3.2×10^{-19}	DDD	Yes
GNAI1	7 (1)	1	2.1×10^{-13}	DDD	No
CSNK2A1	7	0	1.4×10^{-12}	DDD	Yes
PPM1D	0	5 (1)	6.3×10^{-12}	Meta	No
CNOT3	5	2 (1)	5.2×10^{-11}	DDD	Yes
MSL3	0	4	2.2×10^{-10}	DDD	No
KCNQ3	4 (3)	0	3.4×10^{-10}	Meta	Yes
ZBTB18	1 (1)	4	1.4×10^{-9}	DDD	No
PUF60	4 (1)	3	2.6×10^{-9}	DDD	No
TCF20	1	5	2.7×10^{-9}	DDD	No
KMT5B	0 (2)	2 (3)	2.9×10^{-9}	Meta	No
CHD4	8 (1)	1	7.6×10^{-9}	DDD	No
SET	0	3	1.2×10^{-7}	DDD	No
QRICH1	0	3 (1)	3.6×10^{-7}	Meta	No

The numbers of unrelated individuals with independent DNMs are given for protein-truncating variants (PTV) and missense variants. Counts of individuals in other cohorts are given in brackets if present. The *P* value reported is the minimum *P* value from the testing of the DDD dataset or the meta-analysis dataset. The subset providing the *P* value is also listed. Mutations are considered clustered if the *P* value from proximity clustering of DNMs is less than 0.01.

DNMs in known seizure-associated genes, 56% of individuals had no coded terms related to seizures and/or epilepsy. These findings suggest that the power of increased sample size far outweighs specific phenotypic expressivity owing to the shared genetic aetiology between individuals with and without epilepsy in our cohort. Despite this, nearly three times as many individuals with seizures had a DNM in a seizure-associated gene compared to individuals without seizures (Extended Data Fig. 5c). With matched sample sizes, more genes exceeded genome-wide significance in seizure samples than in unstratified samples (Extended Data Fig. 5d). This highlights the cost–benefit effect of recruiting a phenotypically more homogenous cohort.

The large number of genes with genome-wide significance identified in the analyses above allows us to compare empirically different experimental strategies for novel gene discovery in a genetically heterogeneous cohort. We compared the power of exome and genome sequencing to detect genes with genome-wide significance, assuming that budget and not samples are limiting, under different scenarios of cost ratios and sensitivity ratios (Extended Data Fig. 6a). At current cost ratios (exome analysis costs 30–40% of genome analysis) and with a plausible sensitivity differential (genome analysis detects 5% more exonic variants than exome analysis[29]), exome sequencing detects more than twice as many genome-wide significant genes. These empirical estimates were consistent with power simulations for identifying dominant loss-of-function genes (Extended Data Fig. 6b). In summary, although genome sequencing provides the greatest sensitivity to detect pathogenic variation in a single individual (or outside of the coding region), exome sequencing is more powerful for novel gene discovery for disease (and, analogously, probably currently delivers a lower cost per diagnosis).

Our previous simulations suggested that analysis of a cohort of 4,293 DDD families should be able to detect approximately half of all haploinsufficient DD-associated genes at genome-wide significance[15]. Empirically, we have identified 47% (50 out of 107) of haploinsufficient genes that have been previously robustly associated with neurodevelopmental disorders[18]. We hypothesized that genetic testing before recruitment into our study may have depleted the cohort of the most clinically recognizable disorders. Indeed, we observed that the genes associated with the most clinically recognizable disorders were associated with a significant, threefold lower enrichment of truncating DNMs than other DD-associated genes ($P = 8.9 \times 10^{-20}$; approximately 40-fold enrichment for the most clinically recognizable disorders compared to around 120-fold enrichment for cryptic disorders; Fig. 3a). Removing these most recognizable disorders from the analysis, we identified 55% (42 out of 76) of the remaining haploinsufficient DD-associated genes. The known DD-associated haploinsufficient genes that did not reach genome-wide significance were clearly enriched in those with lower mutability, which we would expect to lower the power to detect for these analyses. We identified DD-associated genes (for example, *NRXN2*) with high mutability, low clinical recognizability and yet no signal of enrichment for DNMs in our cohort, as assessed by Δ_{AIC} (the difference between the Akaike's Information Criterion of model 1 and model 2) (Extended Data Fig. 7 and Supplementary Table 4). The current analyses call into question whether these genes really are associated with haploinsufficient neurodevelopmental disorders and highlight the potential for well-powered, gene-discovery analyses to refute previous credence in disease gene associations or previous inferences of an underlying haploinsufficient mechanism.

We estimated the prevalence of pathogenic missense and truncating DNMs within our cohort by increasing the stringency of called DNMs until the observed synonymous DNMs equated that expected under the null-mutation model (Extended Data Fig. 8a), and then quantifying the excess of observed missense and truncating DNMs across all genes (Fig. 3b). We observed an excess of 576 truncating and 1,220 missense mutations, suggesting that 41.8% (1,796 out of 4,293) of the cohort have a pathogenic DNM. This estimate of the number of excess missense and truncating DNMs in our cohort is robust to varying the stringency of DNM calling (Extended Data Fig. 8b). The vast majority of synonymous DNMs are probably benign, as shown by the uniform distribution (Fig. 3d) among genes, irrespective of their tolerance for truncating variation in the general population (as quantified by the 'probability of being loss-of-function intolerant' (P_{LI}) metric[30]). By contrast, missense and truncating DNMs are significantly enriched in genes with the highest probabilities of being intolerant to truncating variation (missense, $P = 1.1 \times 10^{-47}$; truncating, $P = 3.3 \times 10^{-85}$;

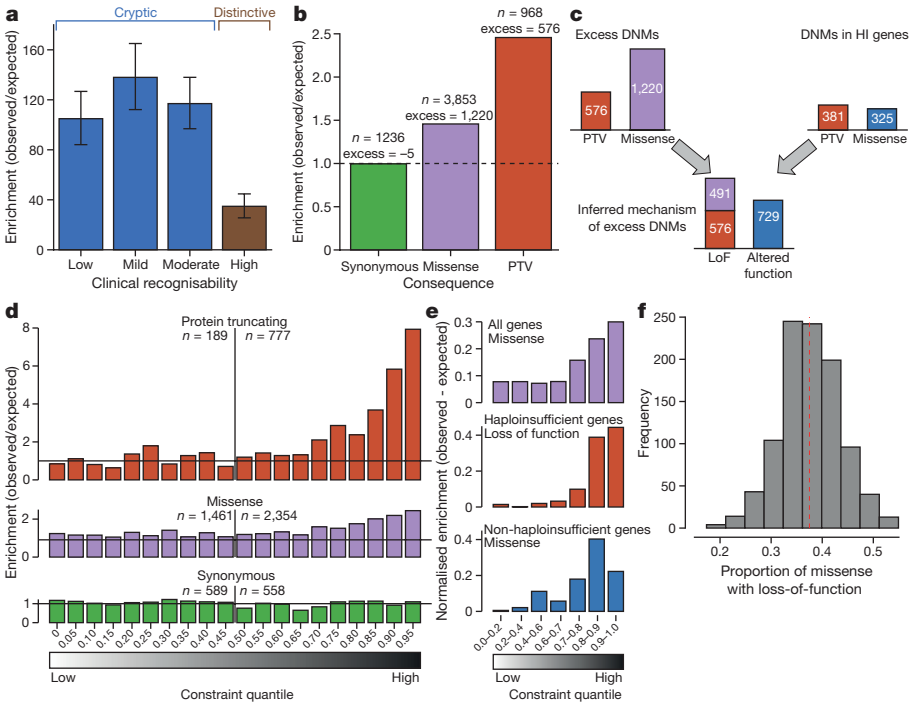


Figure 3 | Excess of DNMs. a, Enrichment ratios of observed to expected loss-of-function DNMs by clinical recognizability for dominant haploinsufficient neurodevelopmental genes as judged by two consultant clinical geneticists. Error bars, 95% CI. **b**, Enrichment of DNMs by consequence normalized relative to the number of synonymous DNMs. **c**, Proportion of excess DNMs with loss-of-function (LoF) or altered-function mechanisms. Proportions are derived from numbers of excess DNMs by consequence, and numbers of excess protein-truncating (PTV) and missense DNMs in dominant haploinsufficient (HI) genes. **d**, Enrichment ratios of observed to expected DNMs by P_{LI}-constraint quantile for protein-truncating, missense and synonymous DNMs. Counts of DNMs in each lower and upper half of the quantiles are provided. Vertical lines divide the lower and upper halves of the constraint scale. Horizontal lines indicate where the observed to expected ratio equals 1. **e**, Normalized excess of observed to expected DNMs by P_{LI}-constraint quantile. This includes missense DNMs within all genes (top), loss of function includes missense DNMs in dominant haploinsufficient genes (middle) and missense DNMs in dominant non-haploinsufficient genes (genes with dominant-negative or activating mechanisms, bottom). **f**, Proportion of excess missense DNMs with a loss-of-function mechanism. The red dashed line indicates the proportion in observed excess DNMs at the optimal goodness of fit. The histogram shows the frequencies of estimated proportions from 1,000 permutations, assuming the observed proportion is correct.

Fig. 3d). The P_{LI}-based distributions were similar to distributions that used functional constraint[31] (Extended Data Fig. 9). Only 51% (923 out of 1,796) of these excess missense and truncating DNMs are located in DD-associated dominant genes, with the remainder probably affecting genes not yet associated with DDs. A much higher proportion of the excess truncating DNMs (71%) than missense DNMs (42%) affected known DD-associated genes. This suggests that whereas most haploinsufficient DD-associated genes have already been identified, many DD-associated genes characterized by pathogenic missense DNMs remain to be discovered.

Understanding the mechanism of action of a monogenic disorder is an important prerequisite for designing therapeutic strategies[32]. We tried to estimate the relative proportion of altered-function and loss-of-function mechanisms among the excess DNMs in our cohort, by assuming that the vast majority of truncating mutations operate by a loss-of-function mechanism and by using two independent approaches to estimate the relative contribution of the two mechanisms among the excess missense DNMs (see Methods). First, we used the observed ratio of truncating and missense DNMs within haploinsufficient DD-associated genes to estimate the proportion of the excess missense DNMs that probably act by loss of function (Fig. 3c). This approach estimated that 59% (95% CI = 55–64%) of the excess missense and truncating DNMs operate by loss of function, and 41% by altered function. Second, we took advantage of the different population genetic characteristics of known altered-function and loss-of-function DD-associated genes. Specifically, we observed that these two classes of DD-associated genes are differentially depleted of truncating variation in individuals without overt DDs (P_{LI} metric[30]). We modelled the observed P_{LI} distribution of excess missense DNMs as a mixture of the P_{LI} distributions of known altered-function and loss-of-function DD-associated genes (Fig. 3e, f), and estimated that 63% (95% CI = 50-76%) of excess missense DNMs probably act by altered-function mechanisms. Incorporating the truncating DNMs that cause a loss-of-function mechanism, this approach estimated that 57% (95% CI = 48-66%) of excess missense and truncating DNMs operate by loss of function and 43% by altered function.

We estimated the birth prevalence of monoallelic DDs by using the germline-mutation model to calculate the expected cumulative germline-mutation rate of truncating DNMs in haploinsufficient DD-associated genes and scaling this upwards on the basis of the composition of excess DNMs in the DDD cohort described above (see Methods), correcting for disorders that were under represented in our cohort as a result of previous genetic testing (for example, clinically recognizable disorders and large pathogenic copy-number variations identified by previous chromosomal microarray analysis). This gives a mean prevalence estimate of 0.34% (95% CI = 0.31-0.37), or 1 in 295 births. By factoring in the paternal and maternal age effects on the mutation rate (Fig. 1), we modelled age-specific estimates of birth prevalence (Fig. 4) that range from 1 in 448 (both mother and father aged 20) to 1 in 213 (both mother and father aged 45). Assuming a yearly global birth rate of 18.6 live births per 1,000 individuals, and a mean age when giving birth of 26.6 years, nearly 400,000 of the 140 million annual births will have a DD caused by a DNM.

In summary, we have shown that DNMs account for approximately half of the genetic architecture of severe DDs, and are split roughly equally between loss of function and altered function. Whereas most haploinsufficient DD-associated genes have already been identified, many activating and dominant-negative DD-associated genes have not

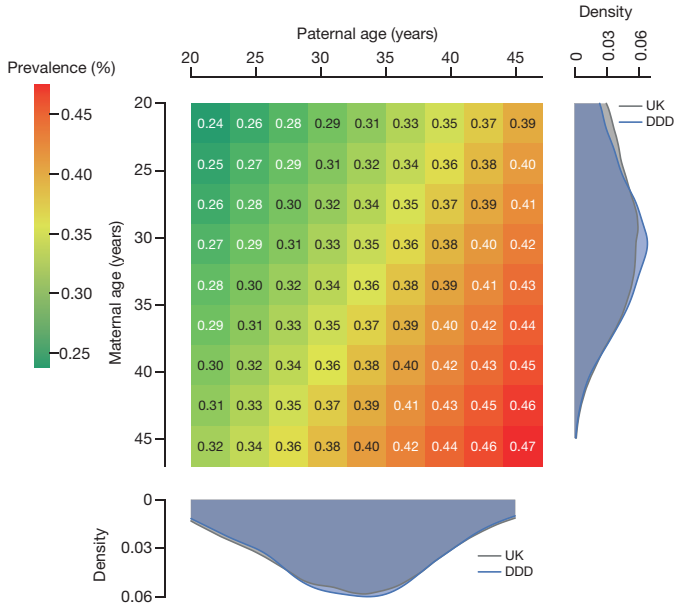


Figure 4 | Prevalence of live births with DDs caused by dominant DNMs. The prevalence within the general population is provided as percentage for combinations of parental ages, extrapolated from the maternal and paternal rates of DNMs. Distributions of parental ages within the DDD cohort and the UK population are shown at the matching parental axis.

yet been described. This probably results from these disorders being individually rarer, being caused by a relatively small number of missense mutations within each gene. It would be valuable to estimate the penetrance of DNMs in the genes we identified as exceeding genome-wide significance, but we cannot formally assess penetrance with our data. Future evaluations could integrate depletion of damaging variation in large healthy populations with patterns of segregation in affected families. Discovery of the remaining dominant DDs requires larger studies and novel, more powerful, analytical strategies for disease-gene association that leverage gene-specific patterns of population variation, specifically the observed depletion of damaging variation. The integration of accurate and complete, quantitative and categorical phenotypic data into the analysis will improve the power to identify ultrarare DDs with distinctive clinical presentations. We have estimated the mean birth prevalence of dominant monogenic DDs to be around 1 in 295, which is greater than the combined impact of trisomies 13, 18 and 21 (ref. 33) and highlights the cumulative population morbidity and mortality imposed by these individually rare disorders.

Note added in proof, other recently published studies have also identified DD-associations for several genes described here, namely *CDK13* (ref. 34), *CHD4* (ref. 34), *CSNK2A1* (ref. 35), *MSL3* (ref. 36), *PPM1D* (ref. 37), *TCF20* (ref. 37) and *ZBTB18* (ref. 38).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 April; accepted 15 December 2016.
Published online 25 January 2017.

1. Sheridan, E. *et al.* Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. *Lancet* **382,** 1350–1359 (2013).
2. Ropers, H. H. Genetics of early onset cognitive impairment. *Annu. Rev. Genomics Hum. Genet.* **11,** 161–187 (2010).
3. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).
4. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).
5. Epi4K Consortium & Epilepsy Phenome/Genome Project. *De novo* mutations in epileptic encephalopathies. *Nature* **501,** 217–221 (2013).
6. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project & Epi4K Consortium. *De novo* mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95,** 360–370 (2014).
7. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
8. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).
9. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).
10. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
11. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
12. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).
13. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
14. Zaidi, S. *et al. De novo* mutations in histone-modifying genes in congenital heart disease. *Nature* **498,** 220–223 (2013).
15. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519,** 223–228 (2015).
16. de Ligt, J., Veltman, J. A. & Vissers, L. E. L. M. Point mutations as a source of *de novo* genetic disease. *Curr. Opin. Genet. Dev.* **23,** 257–263 (2013).
17. Wilkie, A. O. The molecular basis of genetic dominance. *J. Med. Genet.* **31,** 89–98 (1994).
18. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385,** 1305–1314 (2014).
19. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94,** 415–425 (2014).
20. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
21. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48,** 126–133 (2016).
22. Wong, W. S. *et al.* New observations on maternal age effect on germline *de novo* mutations. *Nat. Commun.* **7,** 10486 (2016).
23. Samocha, K. E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
24. Ferry, Q. *et al.* Diagnostically relevant facial gestalt information from ordinary photos. *eLife* **3,** e02020 (2014).
25. Hirata, H. *et al. ZC4H2* mutations are associated with arthrogryposis multiplex congenita and intellectual disability through impairment of central and peripheral synaptic plasticity. *Am. J. Hum. Genet.* **92,** 681–695 (2013).
26. Homan, C. C. *et al.* Mutations in *USP9X* are associated with X-linked intellectual disability and disrupt neuronal cell migration and growth. *Am. J. Hum. Genet.* **94,** 470–478 (2014).
27. Liu, J. *et al. SMC1A* expression and mechanism of pathogenicity in probands with X-linked Cornelia de Lange syndrome. *Hum. Mutat.* **30,** 1535–1542 (2009).
28. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet.* **47,** 1363–1369 (2015).
29. Meynert, A. M., Ansari, M., FitzPatrick, D. R. & Taylor, M. S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15,** 247 (2014).
30. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
31. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).
32. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14,** 681–691 (2013).
33. Springett, A. *et al. Congenital Anomaly Statistics 2011: England and Wales.* (2013).
34. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48,** 1060–1065 (2016).
35. Okur, V. *et al. De novo* mutations in *CSNK2A1* are associated with neurodevelopmental abnormalities and dysmorphic features. *Hum. Genet.* **135,** 699–705 (2016).
36. El Chehadeh, S. *et al.* Dominant variants in the splicing factor *PUF60* cause a recognizable syndrome with intellectual disability, heart defects and short stature. *Eur. J. Hum. Genet.* **25,** 43–51 (2016).
37. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19,** 1194–1196 (2016).
38. Cohen, J. S. *et al.* Further evidence that *de novo* missense and truncating variants in *ZBTB18* cause intellectual disability with variable features. *Clin. Genet.* http://dx.doi.org/10.1111/cge.12861 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the families for their participation and patience. We are grateful to the Exome Aggregation Consortium for making their data available. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute (grant WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). The research team acknowledges the support of the National Institutes for Health Research, through the Comprehensive Clinical Research Network. We thank the Sanger Human Genome Informatics team, the Sample Management team, the Illumina High-Throughput team, the New Pipeline Group team, the DNA pipelines team and the Core Sequencing team for their support in generating and processing the data. D.R.F. is funded through an MRC Human Genetics Unit program grant to the University of Edinburgh. Finally we acknowledge the contribution of two esteemed DDD clinical collaborators, J. Tolmie and L. Brueton, who died during the course of the study.

Author Contributions Patient recruitment and phenotyping: M.Ah., U.A., H.A., R.A., M.Ba., S.Ba., D.Bar., A.Ba., P.B., D.Bat., C.Be., J.Be., B.B., M.B.-G., E.B., M.Bl., D.Boh., L.Bo., D.Bou., L.Br., A.Br., C.Br., K.B., D.J.B., J.Bu., N.Ca., B.C., K.C., D.C., A.Cl., S.Clas., J.C.-S., V.C., A.Coa., T.C., A.Col., M.N.C., F.C., N.Co., H.C., L.C., G.C., Y.C., M.D., T.D., R.D., S.Da., J.D., C.De., G.D., A.Di., A.Dob., A.Don., D.Donna., D.Donne., C.Do., A.Dou., S.Do., A.Du., J.E., S.El., I.E., F.E., K.E., S.Ev., T.F., R.F., F.F., N.F., A.Fry, A.Frye., C.G., L.Ga., N.G., R.G., H.G., J.G., D.G., A.G., P.G., L.Gr., R.Har., L.Ha., V.H., R.Haw., S.Hel., A.H., S.Hew., E.H., S.Holden, M.Ho., S.Holder, G.H., T.H., M.Hu., J.H., S.I., M.I., L.I., A.J., J.J., L.J., D.Joh., E.J., D.Jos., S.J., B.Ka., S.K., B.Ke., H.K., U.K., E.Kin., G.K., C.K., E.Kiv., A.K., D.Ku., V.K.A.K., K.L., W.L., A.L., C.La., M.L., D.L., C.Lo., G.L., S.A.L., A.Mag., E.Ma., A.Mal., S.Ma., K.Mark., K.Mart., U.M., E.Mc., V.Mc., M.M., R.M., K.Mc., S.McK., D.J.M., S.McN., C.M., S.Me., K.Me., Z.M., A.Mi., E.Mi., S.Moh., T.M., D.M., S.Mor., J.M., H.Mug., V.Mu., H.Mur., S.N., A.Ne., L.N., R.N.-E., A.No., R.O., C.O., K.-R.O., S.-M.P., M. J.P., C.Pa., J.Pa., S.Pa., J.Ph., D.T.P., C.Po., J.Po., N.P., K.P., S.Pr., A.Pri., A.Pro., H.P., O.Q., N.R., J.Rank., L.Ra., D.Ri., L.Ro., E.Rob., J.Ro., P.R., G.R., A.R., E.Ros., A.Sag., S.Sa., J.S., R.Sa., A.Sar., S.Sc., R.Sc., I.Sc., A.Selb., A.Sell., C.S., N.S., S.Sh., C.S.-S., E.Shea., D.S., E.Sher., I.Si., R.Si., Z.S., A.Sm., K.S., S.Sm., L.S., M.Sp., M.Sq., F.S., H.S., V.St., M.Su., V.Su., E.Sw., K.T.-B., C.Ta., R.T., M.Tein, I.K.T., J.T., M.Ti., S.T., A.T., B.T., C.Tu., P.T., C.Ty., A.V., V.V., P.Va., J.V., E.Wa., S.Wa., J.W., A.W., D.We., M.Wh., S.Wil., D.Wi., N.W., L.W., G.W., C.W., M.Wr., L.Y., M.Y., H.V.F. and D.R.F. Sample and data processing: S.Clay., T.W.F., E.P., D.Ra., K.A., D.M.B., T.B., P.J., N.K., L.E.M., A.R.T., A.P.B., S.Br., E.C., I.C., E.G., S.G., L.Hi., B.H., R.K., D.P., M.Po., J.Rand., G.J.S., S.Wid. and E.Wi. Validation experiments: J.F.M., E.P., D.Ra., A.Si., N.K. and C.F.W. Study design: M.Pa., H.V.F., C.F.W., D.R.F., J.C.B. and M.E.H. Method development and data analysis: J.F.M., S.Clay., T.W.F., J.K., E.P., D.Ra., A.Si., S.A., N.A., M.Al., P.J., W.D.J., D.Ki., T.S., J.A., D.d.V., L.He, R.R., G.J.S., P.Vi., C.N., H.V.F., C.F.W., D.R.F., J.C.B. and M.E.H. Data interpretation: J.F.M., H.V.F., C.F.W., D.R.F., J.C.B. and M.E.H. Writing: J.F.M., C.F.W., D.R.F. and M.E.H. Experimental and analytical supervision: M.Pa., H.V.F., C.F.W., D.R.F., J.C.B. and M.E.H. Project Supervision: M.E.H.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.E.H. (meh@sanger.ac.uk).

Reviewer Information *Nature* thanks D. Goldstein, B. Neale and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Jeremy F. McRae[1], Stephen Clayton[1], Tomas W. Fitzgerald[1], Joanna Kaplanis[1], Elena Prigmore[1], Diana Rajan[1], Alejandro Sifrim[1], Stuart Aitken[2], Nadia Akawi[1], Mohsan Alvi[3], Kirsty Ambridge[1], Daniel M. Barrett[1], Tanya Bayzetinova[1], Philip Jones[1], Wendy D. Jones[1], Daniel King[1], Netravathi Krishnappa[1], Laura E. Mason[1], Tarjinder Singh[1], Adrian R. Tivey[1], Munaza Ahmed[4,5,6], Uruj Anjum[7], Hayley Archer[8,9], Ruth Armstrong[10], Jana Awada[1], Meena Balasubramanian[11], Siddharth Banka[12], Diana Baralle[4,5,6], Angela Barnicoat[13], Paul Batstone[14], David Baty[15], Chris Bennett[16], Jonathan Berg[15], Birgitta Bernhard[17], A. Paul Bevan[1], Maria Bitner-Glindzicz[13], Edward Blair[18], Moira Blyth[16], David Bohanna[19], Louise Bourdon[17], David Bourn[20], Lisa Bradley[21], Angela Brady[17], Simon Brent[1], Carole Brewer[22], Kate Brunstrom[13], David J. Bunyan[4,5,6], John Burn[20], Natalie Canham[17], Bruce Castle[22], Kate Chandler[12], Elena Chatzimichali[1], Deirdre Cilliers[18], Angus Clarke[8,9], Susan Clasper[18], Jill Clayton-Smith[12], Virginia Clowes[17], Andrea Coates[16], Trevor Cole[19], Irina Colgiu[1], Amanda Collins[4,5,6], Morag N. Collinson[4,5,6], Fiona Connell[23], Nicola Cooper[19], Helen Cox[19], Lara Cresswell[24], Gareth Cross[25], Yanick Crow[12], Mariella D'Alessandro[14], Tabib Dabir[21], Rosemarie Davidson[26], Sally Davies[8,9], Dylan de Vries[1], John Dean[14], Charu Deshpande[23], Gemma Devlin[22], Abhijit Dixit[25], Angus Dobbie[16], Alan Donaldson[27], Dian Donnai[12], Deirdre Donnelly[21], Carina Donnelly[12], Angela Douglas[28], Sofia Douzgou[12], Alexis Duncan[26], Jacqueline Eason[25], Sian Ellard[22], Ian Ellis[28], Frances Elmslie[7], Karenza Evans[8,9], Sarah Everest[22], Tina Fendick[23], Richard Fisher[20], Frances Flinter[23], Nicola Foulds[4,5,6], Andrew Fry[8,9], Alan Fryer[28], Carol Gardiner[26], Lorraine Gaunt[12], Neeti Ghali[17], Richard Gibbons[18], Harinder Gill[29], Judith Goodship[20], David Goudie[15], Emma Gray[1], Andrew Green[29], Philip Greene[2], Lynn Greenhalgh[28], Susan Gribble[1], Rachel Harrison[25], Lucy Harrison[4,5,6], Victoria Harrison[4,5,6], Rose Hawkins[27], Liu He[1], Stephen Hellens[20], Alex Henderson[20], Sarah Hewitt[16], Lucy Hildyard[1], Emma Hobson[16], Simon Holden[10], Muriel Holder[17], Susan Holder[17], Georgina Hollingsworth[13], Tessa Homfray[7], Mervyn Humphreys[21], Jane Hurst[13], Ben Hutton[1], Stuart Ingram[11], Melita Irving[23], Lily Islam[19], Andrew Jackson[2], Joanna Jarvis[19], Lucy Jenkins[13], Diana Johnson[11], Elizabeth Jones[12], Dragana Josifova[23], Shelagh Joss[26], Beckie Kaemba[24], Sandra Kazembe[24], Rosemary Kelsell[1], Bronwyn Kerr[12], Helen Kingston[12], Usha Kini[18], Esther Kinning[26], Gail Kirby[19], Claire Kirk[21], Emma Kivuva[22], Alison Kraus[16], Dhavendra Kumar[8,9], V. K. Ajith Kumar[13], Katherine Lachlan[4,5,6], Wayne Lam[2], Anne Lampe[2], Caroline Langman[23], Melissa Lees[13], Derek Lim[19], Cheryl Longman[26], Gordon Lowther[26], Sally A. Lynch[29], Alex Magee[21], Eddy Maher[2], Alison Male[13], Sahar Mansour[7], Karen Marks[7], Katherine Martin[25], Una Maye[28], Emma McCann[30], Vivienne McConnell[21], Meriel McEntagart[7], Ruth McGowan[14], Kirsten McKay[19], Shane McKee[21], Dominic J. McMullan[19], Susan McNerlan[21], Catherine McWilliam[14], Sarju Mehta[10], Kay Metcalfe[12], Anna Middleton[1], Zosia Miedzybrodzka[14], Emma Miles[12], Shehla Mohammed[23], Tara Montgomery[20], David Moore[2], Sian Morgan[8,9], Jenny Morton[19], Hood Mugalaasi[8,9], Victoria Murday[26], Helen Murphy[12], Swati Naik[19], Andrea Nemeth[18], Louise Nevitt[11], Ruth Newbury-Ecob[27], Andrew Norman[19], Rosie O'Shea[29], Caroline Ogilvie[23], Kai-Ren Ong[19], Soo-Mi Park[10], Michael J. Parker[11], Chirag Patel[19], Joan Paterson[10], Stewart Payne[17], Daniel Perrett[1], Julie Phipps[18], Daniela T. Pilz[26], Martin Pollard[1], Caroline Pottinger[30], Joanna Poulton[18], Norman Pratt[15], Katrina Prescott[16], Sue Price[18], Abigail Pridham[18], Annie Procter[8,9], Hellen Purnell[18], Oliver Quarrell[11], Nicola Ragge[19], Raheleh Rahbari[1], Josh Randall[1], Julia Rankin[22], Lucy Raymond[10], Debbie Rice[15], Leema Robert[23], Eileen Roberts[27], Jonathan Roberts[10], Paul Roberts[16], Gillian Roberts[28], Alison Ross[14], Elisabeth Rosser[13], Anand Saggar[7], Shalaka Samant[14], Julian Sampson[8,9], Richard Sandford[10], Ajoy Sarkar[25], Susann Schweiger[15], Richard Scott[13], Ingrid Scurr[27], Ann Selby[25], Anneke Seller[18], Cheryl Sequeira[17], Nora Shannon[25], Saba Sharif[19], Charles Shaw-Smith[22], Emma Shearing[11], Debbie Shears[18], Eamonn Sheridan[16], Ingrid Simonic[10], Roldan Singzon[17], Zara Skitt[12], Audrey Smith[16], Kath Smith[11], Sarah Smithson[27], Linda Sneddon[20], Miranda Splitt[20], Miranda Squires[16], Fiona Stewart[21], Helen Stewart[18], Volker Straub[20], Mohnish Suri[25], Vivienne Sutton[28], Ganesh Jawahar Swaminathan[1], Elizabeth Sweeney[28], Kate Tatton-Brown[7], Cat Taylor[11], Rohan Taylor[7], Mark Tein[19], I. Karen Temple[4,5,6], Jenny Thomson[16], Marc Tischkowitz[10], Susan Tomkins[27], Audrey Torokwa[4,5,6], Becky Treacy[10], Claire Turner[22], Peter Turnpenny[22], Carolyn Tysoe[22], Anthony Vandersteen[17], Vinod Varghese[8,9], Pradeep Vasudevan[24], Parthiban Vijayarangakannan[1], Julie Vogt[19], Emma Wakeling[17], Sarah Wallwark[10], Jonathon Waters[13], Astrid Weber[28], Diana Wellesley[4,5,6], Margo Whiteford[26], Sara Widaa[1], Sarah Wilcox[10], Emily Wilkinson[1], Denise Williams[19], Nicola Williams[26], Louise Wilson[13], Geoff Woods[10], Christopher Wragg[27], Michael Wright[20], Laura Yates[20], Michael Yau[23], Chris Nellåker[31,32,33], Michael Parker[34], Helen V. Firth[1,10]§, Caroline F. Wright[1]§, David R. FitzPatrick[1,2]§, Jeffrey C. Barrett[1]§ & Matthew E. Hurles[1]§

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [3]Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK. [4]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Coxford Road, Southampton SO16 5YA, UK. [5]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Odstock Road, Salisbury, Wiltshire SP2 8BJ, UK. [6]Faculty of Medicine, University of Southampton, Building 85, Life Sciences Building, Highfield Campus, Southampton SO17 1BJ, UK. [7]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK. [8]Institute of Medical Genetics, University Hospital of Wales, Heath Park, Cardiff CF14 4XW, UK. [9]Department of Clinical Genetics, Block 12, Glan Clwyd Hospital, Rhyl, Denbighshire LL18 5UJ, UK. [10]East Anglian Medical Genetics Service, Box 134, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [11]Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Western Bank, Sheffield S10 2TH, UK. [12]Manchester Centre for Genomic Medicine, St Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL, UK. [13]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, Great Ormond Street, London WC1N 3JH, UK. [14]North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Foresterhill, Aberdeen AB25 2ZD, UK. [15]East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee DD1 9SY, UK. [16]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Chapeltown Road, Leeds LS7 4SA, UK. [17]North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre, Northwick Park and St Mark's NHS Trust Watford Road, Harrow HA1 3UJ, UK. [18]Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, The Churchill Old Road, Oxford OX3 7LJ, UK. [19]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Edgbaston, Birmingham B15 2TG, UK. [20]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK. [21]Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, UK. [22]Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Gladstone Road, Exeter EX1 2ED, UK. [23]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK. [24]Leicestershire Genetics Centre, University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary (NHS Trust), Leicester LE1 5WW, UK. [25]Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, The Gables, Hucknall Road, Nottingham NG5 1PB, UK. [26]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute of Medical Genetics, Yorkhill Hospital, Glasgow G3 8SJ, UK. [27]Bristol Genetics Service (Avon, Somerset, Gloucs and West Wilts), University Hospitals Bristol NHS Foundation Trust, St Michael's Hospital, St Michael's Hill, Bristol BS2 8DT, UK. [28]Merseyside and Cheshire Genetics Service, Liverpool Women's NHS Foundation Trust, Department of Clinical Genetics, Royal Liverpool Children's Hospital Alder Hey, Eaton Road, Liverpool L12 2AP, UK. [29]National Centre for Medical Genetics, Our Lady's Children's Hospital, Crumlin, Dublin 12, Ireland. [30]Department of Clinical Genetics, Block 12, Glan Clwyd Hospital, Rhyl, Denbighshire LL18 5UJ, Wales, UK. [31]Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Level 3, Women's Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK. [32]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Old Road Campus Research Building, Oxford OX3 7DQ, UK. [33]Big Data Institute, University of Oxford, Roosevelt drive, Oxford OX3 7LF, UK. [34]The Ethox Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK.

§These authors jointly supervised this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Family recruitment. At 24 clinical genetics centres within the United Kingdom National Health Service and the Republic of Ireland, 4,293 patients with severe, undiagnosed DDs and their parents (4,125 families) were recruited and systematically phenotyped. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). Families gave informed consent for participation.

Clinical data (growth measurements, family history, developmental milestones, and so on) were collected using a standard restricted-term questionnaire within DECIPHER[39], and detailed developmental phenotypes for the individuals were entered using HPO terms[40]. Saliva samples for the whole family and blood-extracted DNA samples for the probands were collected, processed and quality controlled as previously described[15].

Exome sequencing. Genomic DNA (approximately 1 μg) was fragmented to an average size of 150 base pairs (bp) and a DNA library was created using established Illumina paired-end protocols. Adaptor-ligated libraries were amplified and indexed using polymerase chain reaction (PCR). A portion of each library was used to create an equimolar pool comprising eight indexed libraries. Each pool was hybridized to SureSelect RNA baits (Agilent Human All-Exon V3 Plus with custom ELID C0338371 and Agilent Human All-Exon V5 Plus with custom ELID C0338371) and sequence targets were captured and amplified in accordance with the manufacturer's recommendations. Enriched libraries were analysed by 75-base paired-end sequencing (Illumina HiSeq) following the manufacturer's instructions.

Alignment and calling single-nucleotide variants, insertions and deletions. Mapping of short-read sequences for each sequencing lanelet was carried out using the Burrows-Wheeler aligner (BWA; version 0.59)[41] backtrack algorithm with the GRCh37 1000 Genomes Project phase 2 reference (also known as hs37d5). Sample-level BAM improvement was carried out using the Genome Analysis Toolkit (GATK; version 3.1.1)[42] and SAMtools (version 0.1.19)[43]. This consisted of a realignment of reads around known and discovered indels (insertions and deletions) followed by base quality score recalibration (BQSR), with both steps performed using GATK. Lastly, SAMtools calmd was applied and indexes were created.

Known indels for realignment were taken from the Mills Devine and 1000 Genomes Project Gold set and the 1000 Genomes Project phase low-coverage set, both part of the GATK resource bundle (version 2.2). Known variants for BQSR were taken from dbSNP 137, also part of the GATK resource bundle. Finally, single-nucleotide variants (SNVs) and indels were called using the GATK HaplotypeCaller (version 3.2.2); this was run in multisample calling mode using the complete dataset. GATK Variant Quality Score Recalibration (VQSR) was then computed on the whole dataset and applied to the individual-sample variant calling format (VCF) files. DeNovoGear (version 0.54)[44] was used to detect SNV, insertion and deletion DNMs from child and parental exome data (BAM files).

Variant annotation. Variants in the VCF were annotated with minor allele frequency (MAF) data from a variety of different sources. The MAF annotations used included data from four different populations of the 1000 Genomes Project[45] (American, Asian, African and European), the UK10K cohort, the NHLBI GO Exome Sequencing Project (ESP), the Non-Finnish European (NFE) subset of the Exome Aggregation Consortium (ExAC) and an internal allele frequency generated using unaffected parents from the cohort.

Variants in the VCF were annotated with Ensembl Variant Effect Predictor (VEP)[46] based on Ensembl gene build 76. The transcript with the most severe consequence was selected and all associated VEP annotations were based on the predicted effect of the variant on that particular transcript; where multiple transcripts shared the same most severe consequence, the canonical or longest was selected. We included an additional consequence for variants at the last base of an exon before an intron, where the final base is a guanine, since these variants appear to be as damaging as a splice-donor variant[28]. We categorized variants into three classes by VEP consequence: (1) protein-truncating variants (PTV): splice donor, splice acceptor, stop gained, frameshift, initiator codon and conserved exon terminus variant; (2) missense variants: missense, stop lost, inframe deletion, inframe insertion, coding sequence and protein altering variant; (3) silent variants: synonymous.

DNM filtering. We filtered candidate DNM calls to reduce the false-positive rate but to maximize sensitivity, on the basis of previous results from experimental validation by capillary sequencing of candidate DNMs[15]. Candidate DNMs were excluded if not called by GATK in the child, or called in either parent, or if they had a maximum MAF greater than 0.01. Candidate DNMs were excluded when the forward and reverse coverage differed between reference and alternative alleles, defined as $P < 10^{-3}$ using a Fisher's exact test of coverage from orientation by allele summed across the child and parents.

Candidate DNMs were also excluded if they met two of the three following three criteria: (1) an excess of parental alternative alleles within the cohort at the DNMs position, defined as $P < 10^{-3}$ under a one-sided binomial test given an expected error rate of 0.002 and the cumulative parental depth; (2) an excess of alternative alleles within the cohort in DNMs in a gene, defined as $P < 10^{-3}$ under a one-sided binomial test given an expected error rate of 0.002 and the cumulative depth; or (3) both parents had one or more reads supporting the alternative allele.

If, after filtering, more than one variant was observed in a given gene for a particular trio, only the variant with the highest predicted functional impact was kept (protein truncating > missense > silent).

DNM validation. For candidate DNMs of interest, primers were designed to amplify 150–250-bp products centred around the site of interest. Default primer3 design settings were used with the following adjustments: GC clamp = 1, human mispriming library used. Site-specific primers were tailed with Illumina adaptor sequences. PCR products were generated with JumpStart AccuTaq LA DNA polymerase (Sigma Aldrich), using 40 ng genomic DNA as template. Amplicons were tagged with Illumina PCR primers along with unique barcodes enabling multiplexing of 96 samples. Barcodes were incorporated using Kapa HiFi mastermix (Kapa Biosystems). Samples were pooled and sequenced down one lane of the Illumina MiSeq, using 250 bp paired-end reads. An in-house analysis pipeline extracted the read count per site and classified inheritance status per variant using a maximum likelihood approach (see Supplementary Note).

Individuals with likely pathogenic variants. We previously screened 1,133 individuals for variants that contribute to their disorder[15,18]. All candidate variants in the 1,133 individuals were reviewed by consultant clinical geneticists for relevance to the individuals' phenotypes. Most diagnosable pathogenic variants occurred *de novo* in dominant genes, but a small proportion also occurred in recessive genes or under other inheritance modes. DNMs within dominant DD-associated genes were very probable to be classified as the pathogenic variant for the individuals' disorder. Owing to the time required to review individuals and their candidate variants, we did not conduct a similar review in the remainder of the 4,293 individuals. Instead we defined probable pathogenic variants as candidate DNMs found in autosomal and X-linked dominant DD-associated genes, or candidate DNMs found in hemizygous DD-associated genes in males. 1,136 individuals in the 4,293 cohort had variants either previously classified as pathogenic[15,18], or had a probably pathogenic DNM.

Gene-wise assessment of DNM significance. Gene-specific germline mutation rates for different functional classes were computed[15,23] for the longest transcript in the union of transcripts overlapping the observed DNMs in that gene. We evaluated the gene-specific enrichment of PTV and missense DNMs by computing its statistical significance under a null hypothesis of the expected number of DNMs given the gene-specific mutation rate and the number of considered chromosomes[23].

We also assessed clustering of missense DNMs within genes[15], as expected for DNMs causing activating or dominant-negative mechanisms. We did this by calculating simulated dispersions of the observed number of DNMs within the gene. The probability of simulating a DNM at a specific codon was weighted by the trinucleotide sequence context[15,23]. This allowed us to estimate the probability of the observed degree of clustering given the null model of random mutations.

Fisher's method was used to combine the significance testing of missense + PTV DNM enrichment and missense DNM clustering. We defined a gene as significantly enriched for DNMs if the PTV-enrichment *P* value or the combined missense *P* value was less than 7×10^{-7}, which represents a Bonferroni corrected *P* value of 0.05 adjusted for $4 \times 18{,}500$ tests ($2 \times$ consequence classes tested $\times$ protein coding genes).

Composite face generation. Families were given the option to have photographs of the affected individual(s) uploaded within DECIPHER[39]. Using images of individuals with DNMs in the same gene we generated de-identified realistic average faces (composite faces). Faces were detected using a discriminately trained, deformable-part-model detector[47]. The annotation algorithm identified a set of 36 landmarks per detected face[48] and was trained on a manually annotated dataset of 3,100 images[24]. The average face mesh was created by the Delaunay triangulation of the average constellation of facial landmarks for all patients with a shared genetic disorder.

The averaging algorithm is sensitive to left–right facial asymmetries across multiple patients. For this purpose, we use a template constellation of landmarks based on the average constellations of 2,000 healthy individuals[24]. For each patient, we align the constellation of landmarks to the template with respect to the points along the middle of the face and compute the Euclidean distances between each landmark and its corresponding pair on the template. The faces are mirrored such that the half of the face with the greater difference is always on the same side.

The dataset used for this work may contain multiple photos for one patient. To avoid biasing the average face mesh towards these individuals, we computed an average face for each patient and use these personal averages to compute the final average face. Finally, to avoid any image in the composite dominating owing to variance in illumination between images, we normalized the intensities of pixel values within the face to an average value across all faces in each average. The composite faces were assessed visually to confirm successful ablation of any individually identifiable features. Visual assessment of the composite photograph by two experienced clinical geneticists, alongside the individual patient photos, was performed for all 93 genome-wide significant DD-associated genes for which clinical photos were available for more than one patient, to remove potentially identifiable composite faces as well as quality control on the automated composite face generation process. Eighty-one composite faces were excluded leaving the twelve de-identified composite faces that are shown in Fig. 2 and Extended Data Fig. 3. Each of the twelve composite faces that passed de-identification and quality control was generated from photos of ten or more patients.

Assessing power of incorporating phenotypic information. We previously described a method to assess phenotypic similarity by HPO terms among groups of individuals sharing genetic defects in the same gene[28]. We examined whether incorporating this statistical test improved our ability to identify dominant genes at genome-wide significance. Per gene, we tested the phenotypic similarity of individuals with DNMs in the gene. We combined the phenotypic-similarity *P* value with the genotypic *P* value per gene (the minimum *P* value from the DDD-only and meta-analysis) using Fisher's method. We examined the distribution of differences in *P* value between tests without the phenotypic-similarity *P* value and tests that incorporated the phenotypic-similarity *P* value.

Many individuals (854, 20%) of the DDD cohort experience seizures. We investigated whether testing within the subset of individuals with seizures improved our ability to find associations for seizure-specific genes. A list of 102 seizure-associated genes was curated from three sources: a gene panel for Ohtahara syndrome, a currently used clinical gene panel for epilepsy and a panel derived from DD-associated genes[18]. The *P* values from the seizure subset were compared to *P* values from the complete cohort.

Assessing power of exome versus genome sequencing. We compared the expected power of exome sequencing versus genome sequencing to identify disease genes. Within the DDD cohort, 55 dominant DD-associated genes achieve genome-wide significance when testing for enrichment of DNMs within genes. We did not incorporate missense DNM clustering owing to the large computational requirements for assessing clustering in many replicates.

We assumed a cost of USD$1,000 per individual for genome sequencing. We allowed the cost of exome sequencing to vary relative to genome sequencing, from 10–100%. We calculated the number of trios that could be sequenced under these scenarios. Estimates of the improved power of genome sequencing to detect DNMs in the coding sequence are around 1.05-fold[29] and we increased the number of trios by 1.0–1.2-fold to allow this.

We sampled as many individuals from our cohort as the number of trios and counted which of the 55 DD-associated genes still achieved genome-wide significance for DNM enrichment. We ran 1,000 simulations of each condition and obtained the mean number of genome-wide significant genes for each condition.

Associations with presence of probably pathogenic DNMs. We tested whether phenotypes were associated with the likelihood of having a probably pathogenic DNM. We analysed all collected phenotypes which could be coded in either a binary or quantitative format. Categorical phenotypes (for example, sex coded as male or female) were tested using a Fisher's exact test whereas quantitative phenotypes (for example, duration of gestation coded in weeks) were tested using a logistic regression, using sex as a covariate.

We investigated whether having autozygous regions affected the likelihood of having a diagnostic DNM. Autozygous regions were determined from genotypes in every individual, to obtain the total length per individual. We fitted a logistic regression for the total length of autozygous regions to whether individuals had a probably pathogenic DNM. To illustrate the relationship between length of autozygosity and the occurrence of a probably pathogenic DNM, we grouped the individuals by length and plotted the proportion of individuals in each group with a DNM against the median length of the group.

The effects of parental age on the number of DNMs were assessed using 8,409 high confidence (posterior probability of DNM >0.5) unphased coding and non-coding DNMs in 4,293 individuals. A Poisson multiple regression was fit on the number of DNMs in each individual with both maternal and paternal age at child birth as covariates. The model was fit with the identity link and allowed for over-dispersion. This model used exome-based DNMs, and the analysis was scaled to the whole genome by multiplying the coefficients by a factor of 50, based on approximately 2% of the genome being well covered by our data (exons + introns).

Excess of DNMs by consequence. We identified the threshold for posterior probability of DNM for which the number of observed candidate synonymous DNMs was equal to the number of expected synonymous DNMs. Candidate DNMs with scores below this threshold were excluded. We also examined the probable sensitivity and specificity of this threshold based on validation results for DNMs of a previous publication[15] in which comprehensive experimental validation was performed on 1,133 trios that comprise a subset of the families analysed here.

The numbers of expected DNMs per gene were calculated per consequence from expected mutation rates per gene and the 2,407 male and 1,886 females in the cohort. We calculated the excess of DNMs for missense and PTVs as the ratio of numbers of observed DNMs versus expected DNMs, as well as the difference of observed DNMs minus expected DNMs.

Ascertainment bias within dominant neurodevelopmental genes. We identified 150 autosomal dominant haploinsufficient genes that affect neurodevelopment within our curated DD gene set. Genes affecting neurodevelopment were identified where the affected organs included the brain; or where HPO phenotypes linked to defects in the gene included either an abnormality of brain morphology (HP:0012443) or cognitive impairment (HP:0100543) term.

The 150 genes were classified for ease of clinical recognition of the syndrome from gene defects by two consultant clinical geneticists. Genes were rated from 1 (least recognizable) to 5 (most recognizable). Categories 1 and 2 contained 5 and 22 genes, respectively, and so were combined in later analyses. The remaining categories had more than 33 genes per category. The ratio of observed loss-of-function DNMs to expected loss-of-function DNMs was calculated for each recognizability category, along with 95% CIs from a Poisson distribution given observed counts.

We estimated the likelihood of obtaining the observed number of PTV DNMs under two models. Our first model assumed no haploinsufficiency, and mutation counts were expected to follow baseline mutation rates. Our second model assumed fully penetrant haploinsufficiency, and scaled the baseline PTV-mutation expectations by the observed PTV enrichment in our known haploinsufficient neurodevelopmental genes, stratified by clinical recognizability into low (containing genes with our 'low', 'mild' and 'moderate' labels) and high categories. We calculated the likelihoods of both models per gene as the Poisson probability of obtaining the observed number of PTVs, given the expected mutation rates. We computed the Akaike's Information Criterion for each model and ranked them by the difference between model 1 and model 2 (Δ_{AIC}).

Proportion of DNMs with a loss-of-function mechanism. The observed excess of missense/inframe indel DNMs is composed of a mixture of DNMs with loss-of-function mechanisms and DNMs with altered-function mechanisms. We found that the excess of PTV DNMs within dominant haploinsufficient DD-associated genes had a greater skew towards genes with high intolerance for loss-of-function variants than the excess of missense DNMs in dominant non-haploinsufficient genes. We binned genes by the probability of being loss-of-function intolerant[30] constraint decile and calculated the observed excess of missense DNMs in each bin. We modelled this binned distribution as a two-component mixture with the components representing DNMs with a loss-of-function or altered-function mechanism. We identified the optimal mixing proportion for the loss-of-function and altered-function DNMs from the lowest goodness of fit (from a spline fitted to the sum-of-squares of the differences per decile) to missense/inframe indels in all genes across a range of mixtures.

The excess of DNMs with a loss-of-function mechanism was calculated as the excess of DNMs with a VEP loss-of-function consequence, plus the proportion of the excess of missense DNMs at the optimal mixing proportion.

We independently estimated the proportions for loss of function and altered function. We counted PTV and missense/inframe indel DNMs within dominant haploinsufficient genes to estimate the proportion of excess DNMs with a loss-of-function mechanism, but which were classified as missense/inframe indel. We estimated the proportion of excess DNMs with a loss-of-function mechanism as the PTV excess plus the PTV excess multiplied by the proportion of loss of function classified as missense.

Prevalence of DDs from dominant DNMs. We estimated the birth prevalence of monoallelic DDs by using the germline-mutation model. We calculated the expected cumulative germline-mutation rate of truncating DNMs in 238 haploinsufficient DD-associated genes. We scaled this upwards based on the composition of excess DNMs in the DDD cohort using the ratio of excess DNMs ($n=1,816$) to DNMs within dominant haploinsufficient DD-associated genes ($n=412$). Around 10% of DDs are caused by *de novo* copy-number variations[49,50], which are underrepresented in our cohort as a result of previous genetic testing. If included, the excess DNM in our cohort would increase by 21%, therefore we scaled the prevalence estimate upwards by this factor.

Mothers aged 29.9 and fathers aged 29.5 have children with 77 DNMs per genome on average[21]. We calculated the mean number of DNMs expected under

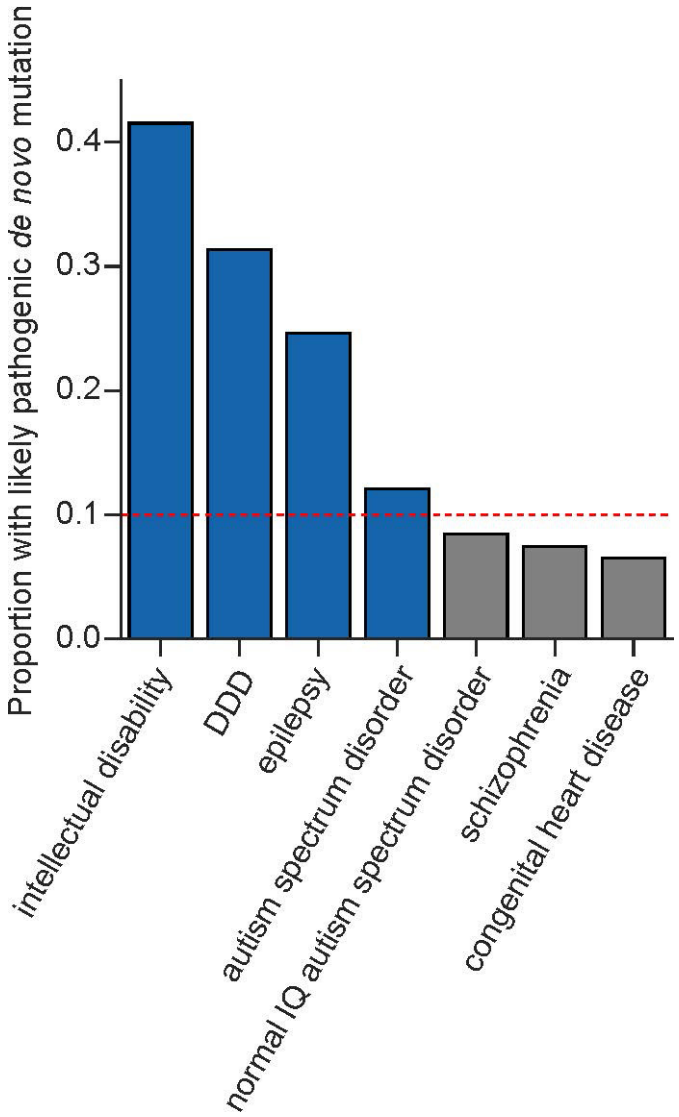
different combinations of parental ages, given our estimates of the extra DNMs per year from older mothers and fathers. We scaled the prevalence to different combinations of parental ages using the ratio of expected mutations at a given age combination to the number expected at the mean cohort parental ages.

To estimate the annual number of live births with DDs caused by DNMs, we obtained country population sizes, birth rates, age at first birth[51], and calculated global birth rate (18.58 live births per 1,000 individuals) and age at first birth (22.62 years), weighted by population size. We calculated the mean age when giving birth (26.57 years) given a total fertility rate of 2.45 children per mother[52], and a mean interpregnancy interval of 29 months[53]. We calculated the number of live births given our estimate of DD prevalence caused by DNMs at this age (0.00288), the global population size (7.4 billion individuals) and the global birth rate.

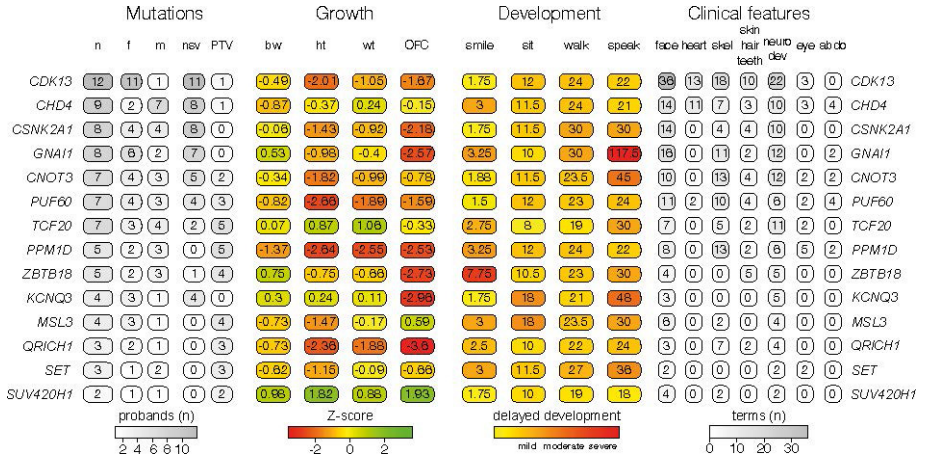
Code availability. Source code for filtering candidate DNMs, testing DNM enrichment, DNM clustering and phenotypic similarity can be found here: https://github.com/jeremymcrae/denovoFilter, https://github.com/jeremymcrae/mupit, https://github.com/jeremymcrae/denovonear and https://github.com/jeremymcrae/hpo_similarity.

Data availability. Exome sequencing and phenotype data are accessible via the European Genome-phenome Archive (EGA) under accession number EGAS00001000775 (https://www.ebi.ac.uk/ega/studies/EGAS00001000775). Details of DD-associated genes are available at www.ebi.ac.uk/gene2phenotype. All other data are available from the corresponding author upon reasonable request.

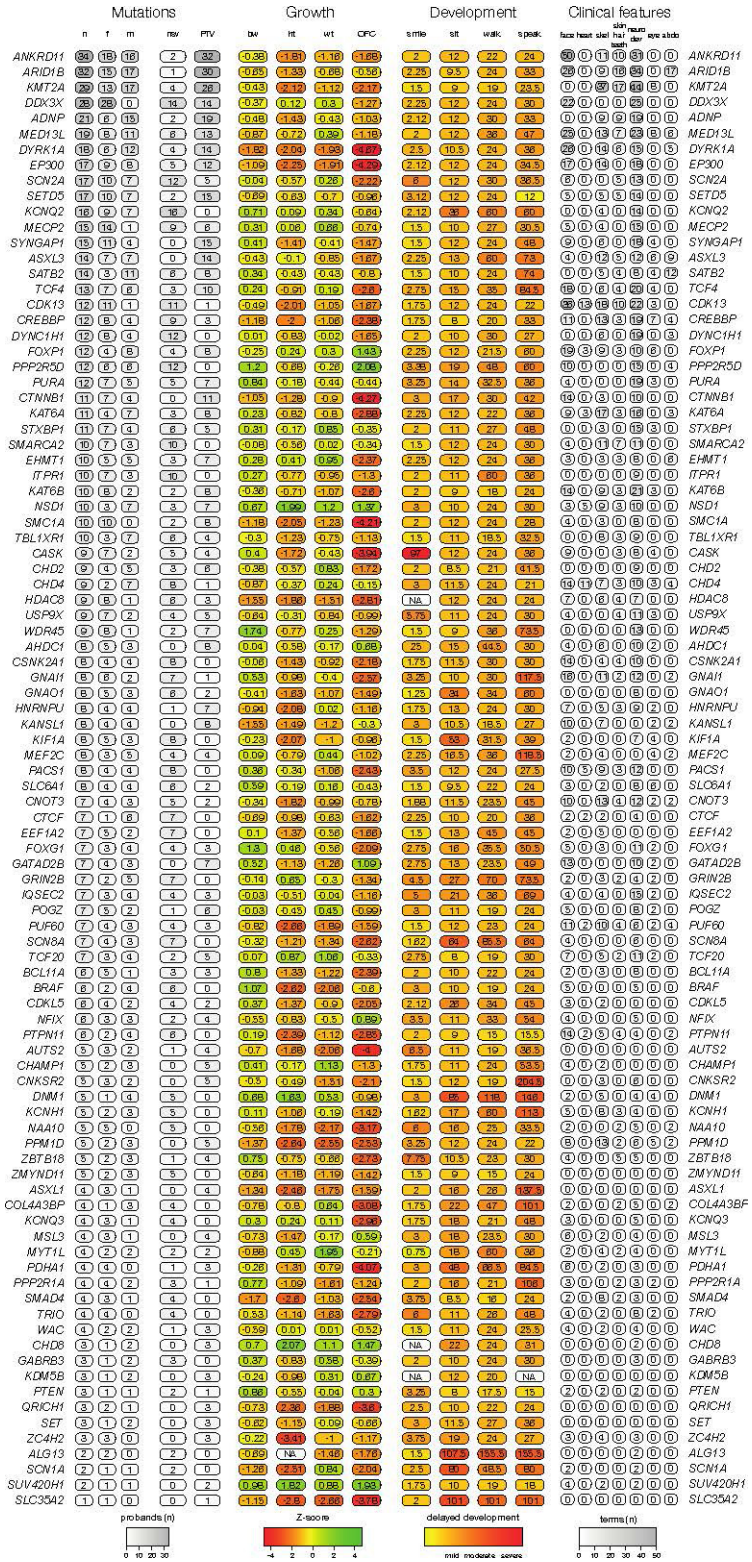
39. Bragin, E. *et al.* DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* **42,** D993–D1000 (2014).
40. Köhler, S. *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85,** 457–464 (2009).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
42. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
44. Ramu, A. *et al.* DeNovoGear: *de novo* indel and point mutation discovery and phasing. *Nat. Methods* **10,** 985–987 (2013).
45. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
46. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
47. Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32,** 1627–1645 (2010).
48. Xiong, X. & De la Torre, F. Supervised Descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 532–539 (Portland, 2013).
49. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43,** 838–846 (2011).
50. Sagoo, G. S. *et al.* Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet. Med.* **11,** 139–146 (2009).
51. Central Intelligence Agency. *The World Factbook.* Vol. 2016 (2016).
52. The World Bank. Fertility rate, total (births per woman). in *World Development Indicators* (2016).
53. Copen, C. E., Thoma, M. E. & Kirmeyer, S. Interpregnancy Intervals in the United States: Data From the Birth Certificate and the National Survey of Family Growth. In *National Vital Statistics Reports* Vol. 64 (National Center for Health Statistics, 2015).



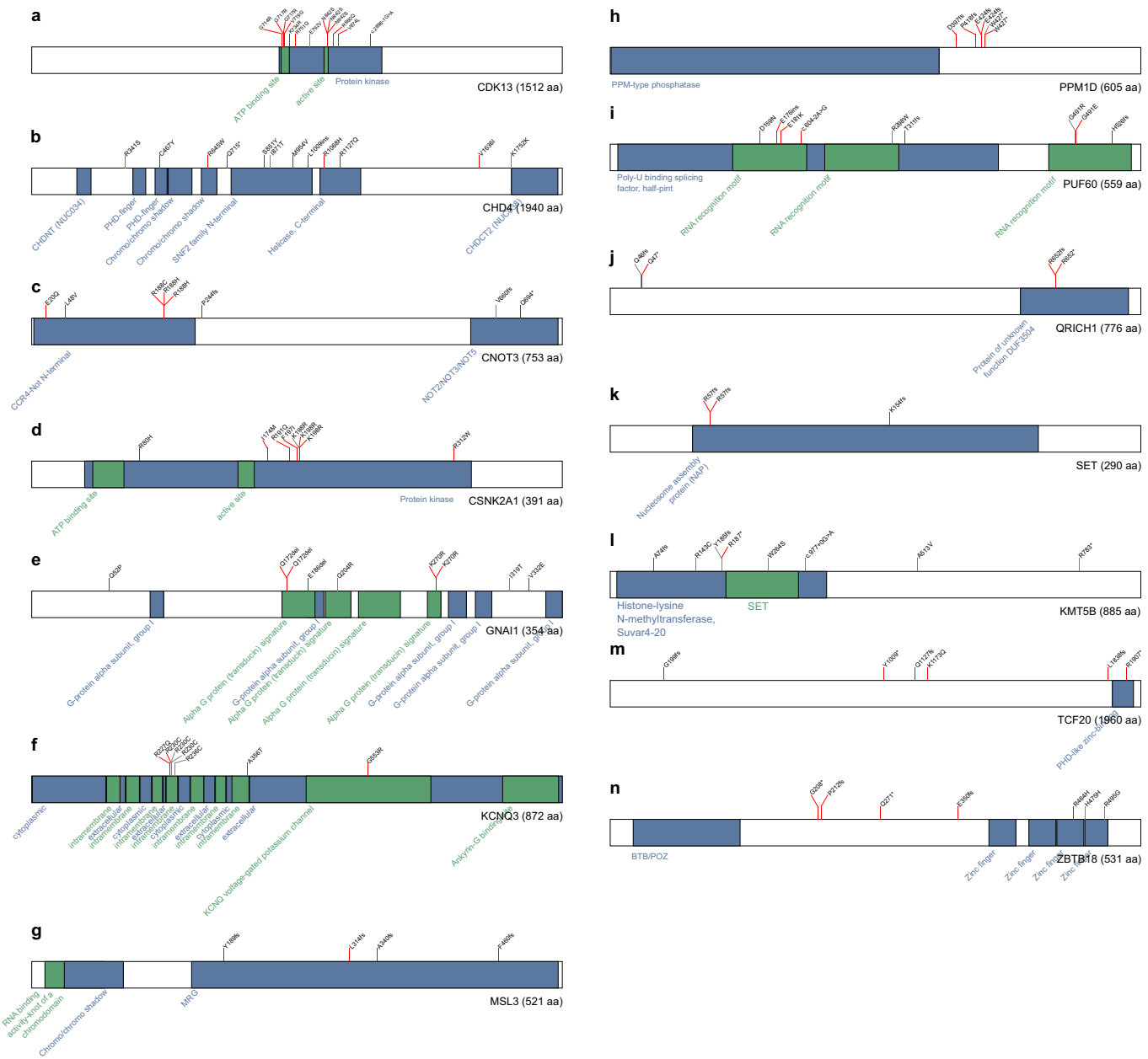
Extended Data Figure 1 | Proportion of individuals with a DNM that is probably pathogenic. Only individuals with protein-altering or protein-truncating DNMs in dominant or X-linked dominant DD-associated genes, or males with DNMs in hemizygous DD-associated genes were included. The proportions given are for those individuals with any DNMs rather than the total number of individuals in each subset. Cohorts included in the DNM meta-analyses are shaded blue.



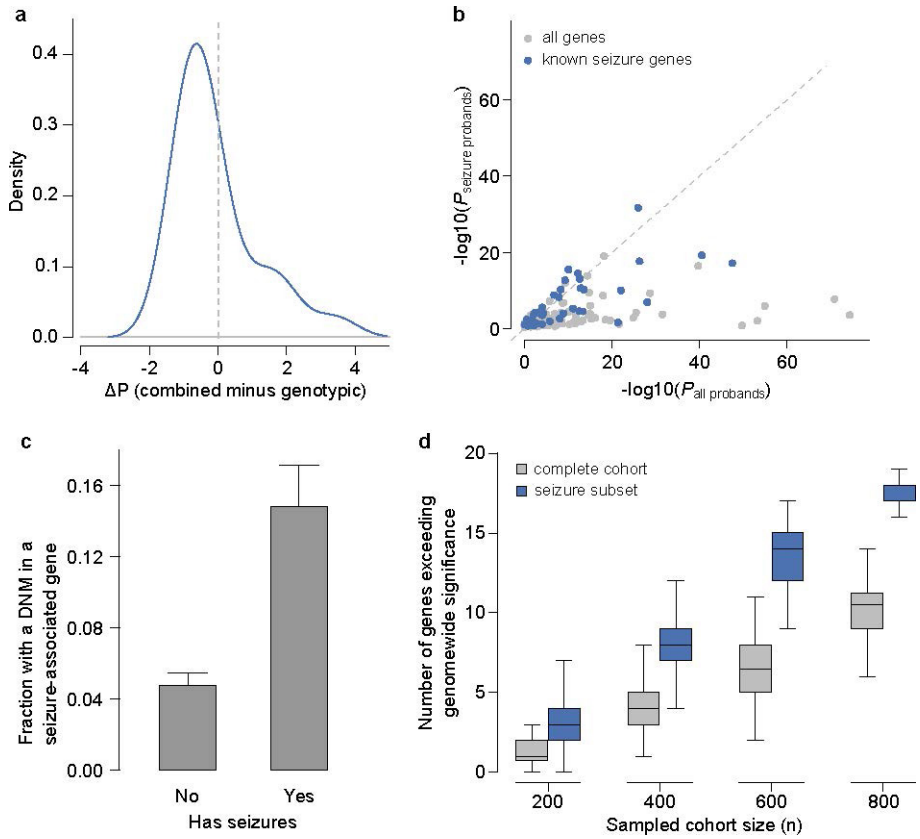
Extended Data Figure 2 | Phenotypic summary of genes without previous compelling evidence. Phenotypes are grouped by type. The first group indicates numbers of individuals with DNMs per gene divided by sex (m, male; f, female), and by functional consequence (NSV, nonsynonymous variant; PTV, protein-truncating variant). The second group indicates mean values for growth parameters: birthweight (bw), height (ht), weight (wt) and occipitofrontal circumference (OFC). Values are given as standard deviations from the healthy population mean derived from ALSPAC (Avon longitudinal study of parents and children) data. The third group indicates the mean age for achieving developmental milestones: age of first social smile, age of first sitting unassisted, age of first walking unassisted and age of first speaking. Values are given in months. The final group summarizes HPO-coded phenotypes per gene, as number of HPO terms within different clinical categories.



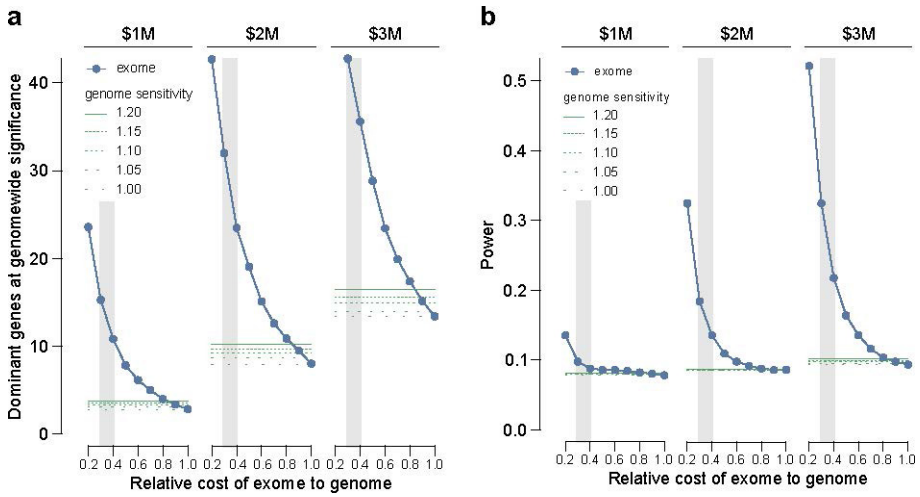
Extended Data Figure 3 | Phenotypic summary of individuals with DNMs in genes achieving genome-wide significance. Phenotypes are grouped by type. The first group indicates numbers of individuals with DNMs per gene divided by sex (m, male; f, female), and by functional consequence (NSV, nonsynonymous variant; PTV, protein-truncating variant). The second group indicates mean values for growth parameters: birthweight (bw), height (ht), weight (wt) and occipitofrontal circumference (OFC). Values are given as standard deviations from the healthy population mean derived from ALSPAC data. The third group indicates the mean age for achieving developmental milestones: age of first social smile, age of first sitting unassisted, age of first walking unassisted and age of first speaking. Values are given in months. The final group summarizes HPO-coded phenotypes per gene, as number of HPO terms within different clinical categories.



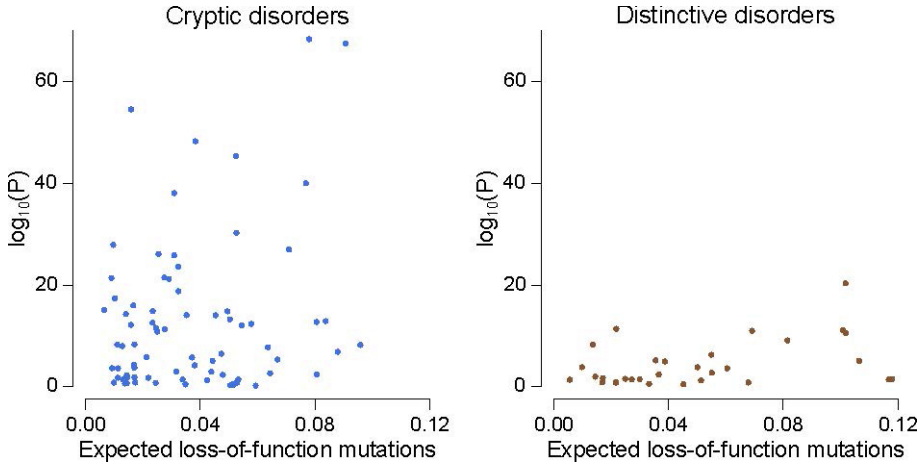
Extended Data Figure 4 | Dispersion of DNMs and domains for each novel gene. **a**, *CDK13*. **b**, *CHD4*. **c**, *CNOT3*. **d**, *CSNK2A1*. **e**, *GNAI1*. **f**, *KCNQ3*. **g**, *MSL3*. **h**, *PPM1D*. **i**, *PUF60*. **j**, *QRICH1*. **k**, *SET*. **l**, *KMT5B*. **m**, *TCF20*. **n**, *ZBTB18*.



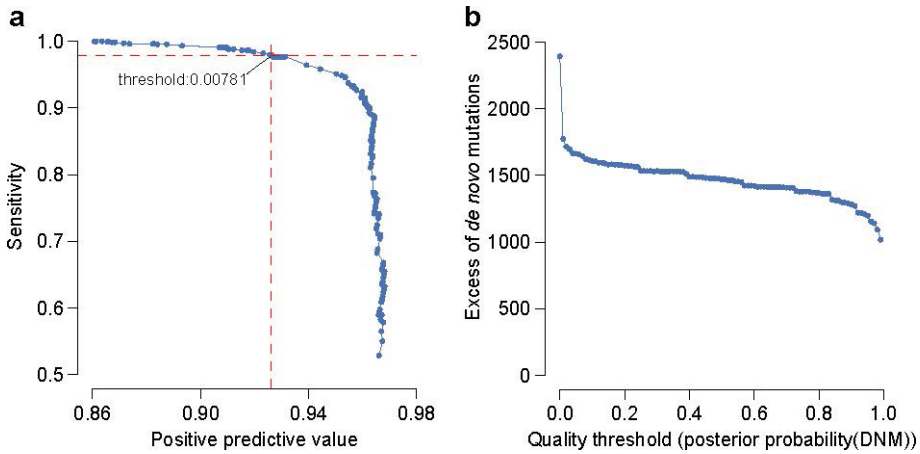
Extended Data Figure 5 | Effect of clustering by phenotype on the ability to identify genome-wide significant genes. a, Comparison of *P* values derived from genotypic information alone versus *P* values that incorporate genotypic information and phenotypic similarity. **b**, Comparison of *P* values from tests in the complete DDD cohort versus tests in the subset with seizures. Genes that were previously linked to seizures are shaded blue. **c**, Proportion of cohort with a DNM in a seizure-associated gene, stratified by seizure-affected status. Error bars, 95% CI. **d**, Comparison of power to identify genome-wide significant genes in probands with seizures, versus the unstratified cohort, at matched sample sizes.



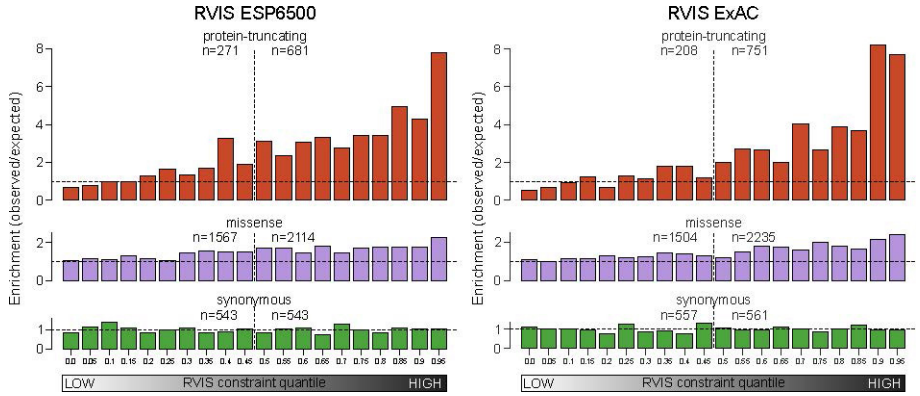
Extended Data Figure 6 | Power of genome versus exome sequencing to discover dominant genes associated with DDs. a, The number of genes exceeding genome-wide significance was estimated at three different fixed budgets ($USD1, 2 or 3 million) and a range of relative sensitivities for genomes versus exomes to detect DNMs. The number of genes identifiable by exome sequencing are shaded blue, whereas the number of genes identifiable by genome sequencing are shaded green. The regions where exome sequencing costs 30–40% of genome sequencing are shaded with a grey background, which corresponds to the price differential in 2016. **b**, Simulated estimates of power to detect loss-of-function genes in the genome at different cohort sizes, given fixed budgets.



Extended Data Figure 7 | Gene-wise significance of neurodevelopmental genes versus the expected number of mutations per gene. Points are shaded by clinical recognizability classification (blue and brown points denote cryptic and distinctive disorders, respectively). Genes have been separated into two plots. Left, genes for cryptic disorders with low, mild or moderate clinical recognizability. Right, genes for distinctive disorders with high clinical recognizability.



Extended Data Figure 8 | Stringency of DNM filtering. a, Sensitivity and specificity of DNM validations within sets filtered using varying thresholds of DNM quality (posterior probability of DNM). The analysed DNMs were restricted to sites identified within the earlier 1,133 trios[15], where all candidate DNMs underwent validation experiments. The labelled value is the quality threshold at which the number of candidate synonymous DNMs equals the number of expected synonymous mutations under a null germline mutation rate. **b**, Excess of missense and loss-of-function DNMs at varying DNM quality thresholds. The DNM excess is adjusted for the sensitivity and specificity at each threshold.



Extended Data Figure 9 | Enrichment of DNMs by consequence type, across functional constraint quantiles for residual variation intolerance scores. A comparison of enrichment for residual variation intolerance score (RVIS) values generated from ESP6500 data (ref. 31) versus ExAC data (obtained from http://genic-intolerance.org/) are provided.

Extended Data Table 1 | Phenotypes tested for association with having a pathogenic DNM

Category	Phenotype	Type	Value	95% CI	P-value
Post-natal	abnormal cranial MRI	Odds ratio	1.365	1.125 – 1.656	0.002
	feeding problems	Odds ratio	1.176	1.01 – 1.369	0.039
	neonatal intensive care	Odds ratio	0.896	0.762 – 1.054	0.190
	anticonvulsant drugs	Odds ratio	0.582	0.246 – 1.377	0.270
Pre-natal	bleeding	Odds ratio	0.892	0.714 – 1.114	0.346
	maternal illness	Odds ratio	0.908	0.764 – 1.079	0.278
	maternal diabetes	Odds ratio	0.787	0.504 – 1.229	0.341
	abnormal scan	Odds ratio	0.839	0.692 – 1.017	0.078
	assisted reproduction	Odds ratio	0.868	0.554 – 1.36	0.584
	increased nuchal translucency	Odds ratio	1.432	0.903 – 2.271	0.126
Family history	consanguinity	Odds ratio	0.234	0.138 – 0.397	8.0×10^{-11}
	similar phenotype parents	Odds ratio	0.295	0.184 – 0.474	5.7×10^{-9}
	similar phenotype relatives	Odds ratio	0.553	0.402 – 0.761	1.5×10^{-4}
	similar phenotype siblings	Odds ratio	0.311	0.23 – 0.421	7.3×10^{-18}
	only patient affected	Odds ratio	2.478	2.001 – 3.068	3.9×10^{-19}
	X-linked inheritance	Odds ratio	0.839	0.436 – 1.613	0.752
	Multiple births	Beta	0.043	-0.058 – 0.144	0.403
	History of pregnancy loss	Beta	-0.039	-0.155 – 0.078	0.516
Developmental milestones	first words	Beta	0.205	0.081 – 0.328	0.001
	walked independently	Beta	0.125	0.016 – 0.235	0.025
	sat independently	Beta	0.050	-0.069 – 0.17	0.408
	social smile	Beta	0.072	-0.066 – 0.211	0.305
Growth	height	Beta	0.008	-0.111 – 0.126	0.897
	birthweight	Beta	-0.018	-0.135 – 0.098	0.756
	OFC	Beta	-0.094	-0.215 – 0.026	0.125
	weight	Beta	-0.331	-1.278 – 0.615	0.493
Age	age at assessment	Beta	0.116	0.015 – 0.217	0.025
	gestation	Beta	0.079	-0.033 – 0.19	0.167
	father's age	Beta	0.137	0.027 – 0.247	0.015
	mother's age	Beta	0.108	-0.003 – 0.219	0.056
Other	phenotypic terms (n)	Beta	0.104	0.004 – 0.203	0.041
	autozygosity length	Beta	-0.185	-0.254 – -0.115	1.6×10^{-7}
	sex (male)	Odds ratio	0.750	0.646 – 0.87	1.6×10^{-4}